BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

(12) UK Patent Application (19) GB (11) 2 346 999 (13) A

(43) Date of A Publication 23.08.2000

(21) Application No 0008337.8

(22) Date of Filing 14.01.2000

(30) Priority Data
(31) 09235952 (32) 22.01.1999 (33) US

(51) INT CL[7]
G10L 15/04 // G10L 101:065

(52) UK CL (Edition R)
G4R RPS R1F
U1S S2105 S2108 S2123 S2127 S2204 S2215 S2322

(56) Documents Cited
GB 2090453 A

(58) Field of Search
UK CL (Edition R) G4R RPS RPW
INT CL[7] G10L 11/00 11/02 15/00 15/04 15/20
Online:WPI, EPODOC, JAPIO

(71) Applicant(s)
Motorola Inc
(Incorporated in USA - Delaware)
Corporate Offices, 1303 East Algonquin Road,
Schaumburg, Illinois 60196, United States of America

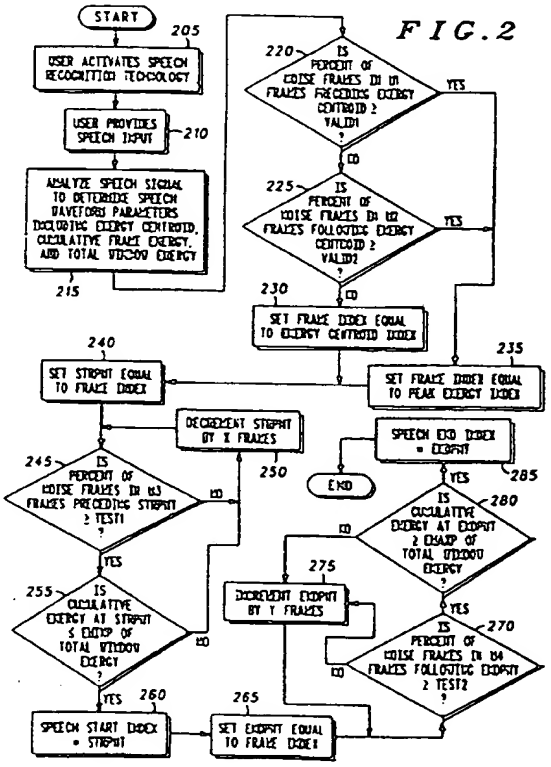
(72) Inventor(s)
William M Kushner
Audrius Polikaitis

(74) Agent and/or Address for Service
Marc Morgan
Motorola Limited, European Intellectual Property
Department, Midpoint,Alencon Link, BASINGSTOKE,
Hampshire, RG21 7PL, United Kingdom

(54) Abstract Title
Communication device for endpointing speech utterances

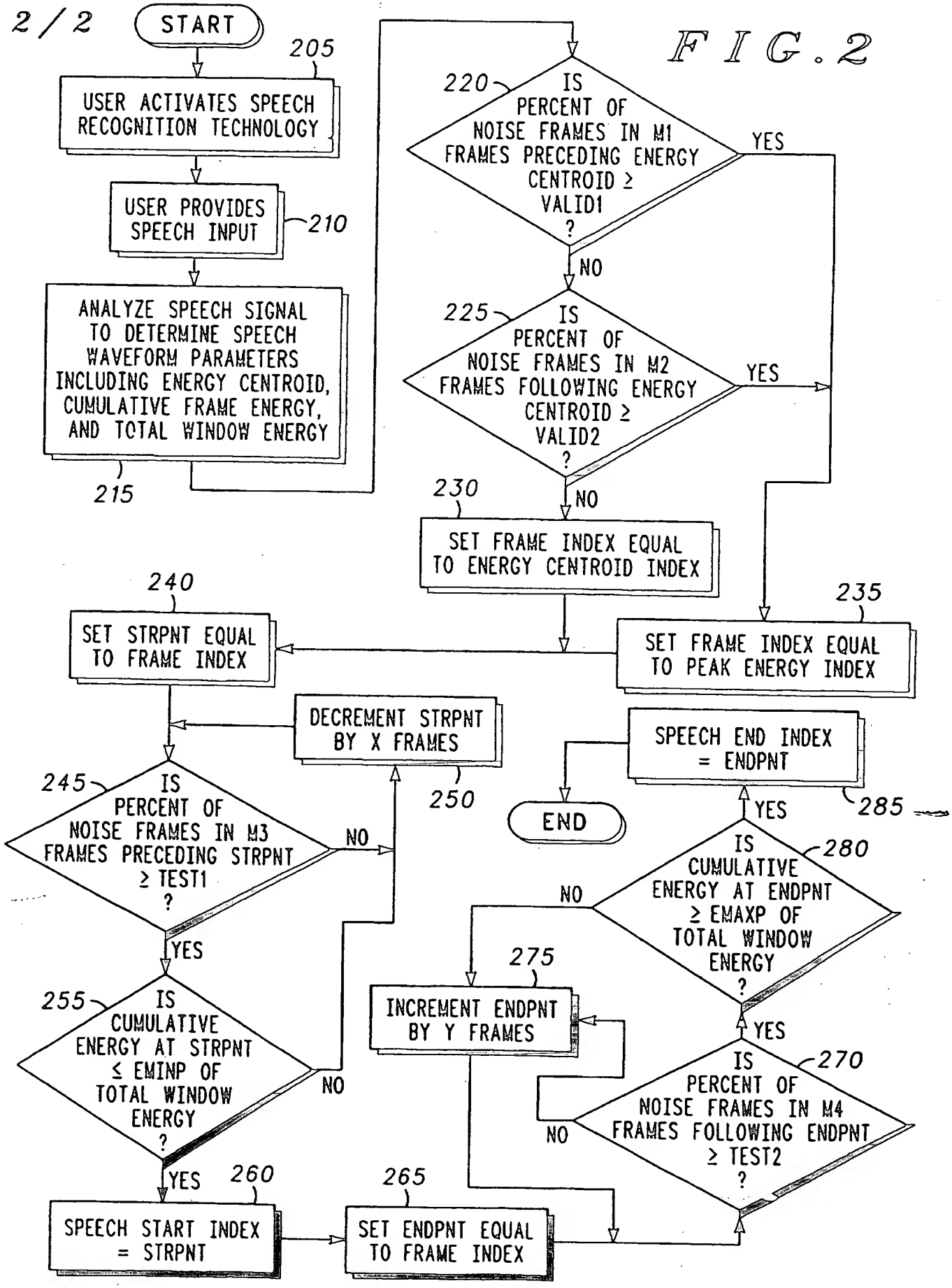
(57) A communication device capable of endpointing speech utterances includes a speech/noise classifier and speech recognition technology. A speech signal is analysed to determine speech waveform parameters within a speech acquisition window 215. The speech waveform parameters are compared to determine the start and end points of the speech utterance. Processing starts at a frame index based on the energy centroid of the speech utterance and analyzes the frames preceding and following the frame index to determine the endpoints. When a potential endpoint is identified, the cumulative nergy is compared to the total energy of the speech acquisition window to determine whether additional speech frames are present 255,280. Accordingly, gaps and pauses in the utterance will not result in an erroneous endpoint determination.

FIG.2



GB 2 346 999 A

FIG. 2


START
205
USER ACTIVATES SPEECH RECOGNITION TECHNOLOGY
USER PROVIDES SPEECH INPUT
210
ANALYZE SPEECH SIGNAL TO DETERMINE SPEECH WAVEFORM PARAMETERS INCLUDING ENERGY CENTROID, CUMULATIVE FRAME ENERGY, AND TOTAL WINDOW ENERGY
215
220
IS PERCENT OF NOISE FRAMES IN M1 FRAMES PRECEDING ENERGY CENTROID ≥ VALID1 ?
YES
NO
225
IS PERCENT OF NOISE FRAMES IN M2 FRAMES FOLLOWING ENERGY CENTROID ≥ VALID2 ?
YES
NO
230
SET FRAME INDEX EQUAL TO ENERGY CENTROID INDEX
235
SET FRAME INDEX EQUAL TO PEAK ENERGY INDEX
240
SET STRPNT EQUAL TO FRAME INDEX
DECREMENT STRPNT BY X FRAMES
250
245
IS PERCENT OF NOISE FRAMES IN M3 FRAMES PRECEDING STRPNT ≥ TEST1 ?
NO
YES
255
IS CUMULATIVE ENERGY AT STRPNT ≤ EMINP OF TOTAL WINDOW ENERGY ?
NO
YES
260
SPEECH START INDEX = STRPNT
265
SET ENDPNT EQUAL TO FRAME INDEX
275
INCREMENT ENDPNT BY Y FRAMES
270
IS PERCENT OF NOISE FRAMES IN M4 FRAMES FOLLOWING ENDPNT ≥ TEST2 ?
NO
YES
280
IS CUMULATIVE ENERGY AT ENDPNT ≥ EMAXP OF TOTAL WINDOW ENERGY ?
NO
YES
285
SPEECH END INDEX = ENDPNT
END

population of expected users are averaged in some manner to create a word model for that word. By averaging speech parameters for the same word spoken by different people, the word model should be usable by most if not all people.

In speaker *dependent* speech recognition devices, the user trains the device by speaking the particular word when prompted by the device. The speech recognition technology then creates a word model based on the input from the user. The speech recognition technology may prompt the user to repeat the word any number of times and then average the speech waveform parameters in some manner to create the word model.

To properly operate speech recognition technology, it is important to consistently identify the start and end endpoints of the speech utterances. Inconsistently identified endpoints may truncate words and may include extraneous noises within the speech waveform acquired by the speech recognition technology. Truncated words and/or noises may result in poorly trained models and cause the speech recognition technology not to work properly when the acquired speech waveform does not match any word model. In addition, truncated words and noises may cause the speech recognition technology to misidentify the acquired speech waveform as another word. In speaker dependent speech recognition devices, problems due to poor endpointing are aggravated when the speech recognition technology permits only a few training utterances.

The prior art describe techniques using threshold energy comparisons, zero crossings analysis, and cross correlation. These methods sequentially analyze speech features from left to right, right to left, or center outwards of the speech waveform. In these techniques, utterances containing pauses or gaps are problematic. Typically, pauses or gaps in an utterance are caused by the nature of the word, the speaking style of the user, and by utterances containing multiple words. Some techniques truncate the word or phrase at the gap, assuming erroneously that the endpoint has been reached. Other techniques use a maximum gap size criteria to combine detected parts of utterances with pauses into a single utterance. In such techniques, a pause longer than a predetermined threshold can cause parts of the utterance to be excluded.

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention is better understood when read in light of the accompanying drawings, in which:

FIG. 1 is a block diagram of a communication device capable of endpointing speech utterances; and

FIG. 2 is a flowchart describing endpointing speech utterances.

DETAILED DESCRIPTION OF THE INVENTION

FIG. 1 is a block diagram of a communication device 100 according to the present invention. Communication device 100 may be a cellular telephone, a portable telephone handset, a two-way radio, a data interface for a computer or personal organizer, or similar electronic device. Communication device 100 includes microprocessor 110 connected to communication interface circuitry 115, memory 120, audio circuitry 130, keypad 140, display 150, and vibrator/buzzer 160.

Microprocessor 110 may be any type of microprocessor including a digital signal processor or other type of digital computing engine. Preferably, microprocessor 110 includes a speech/noise classifier and speech recognition technology. One or more additional microprocessors (not shown) may be used to provide the speech/noise classifier, the speech recognition technology, and the endpointing of the present invention.

Communication interface circuitry 115 is connected to microprocessor 110. The communication interface circuitry is for sending and receiving data. In a cellular telephone, communication interface circuitry 115 would include a transmitter, receiver, and an antenna. In a computer, communication interface circuitry 115 would include a data link to the central processing unit.

Memory 120 may be any type of permanent or temporary memory such as random access memory (RAM), read-only memory (ROM), disk, and other types of

duration and 10 ms are preferred. For each frame, microprocessor 110 determines the frame energy using the following equation:

$$fegy_n = \sum_{i=(n-1)l}^{nl-1} X_i^2 \quad , n = 1, 2, \ldots, N$$

The parameter $fegy_n$ is related to the energy of a frame of sampled data. This can be the actual frame energy or some function of it. X_i are speech samples. I is the number of samples in a data frame, n. N is the total number of frames in the speech acquisition window.

In addition, microprocessor 110 numbers each frame sequentially from 1 through the total number of frames, N. Although the frames may be numbered with the flow (left to right) or against the flow (right to left) of the voice waveform, the frames are preferably numbered with the flow of the waveform. Consequently, each frame has a frame number, n, corresponding to the position of the frame in the speech acquisition window.

Microprocessor 110 has a speech/noise classifier for determining whether each frame is speech or noise. Any speech/noise classifier may be used. However, the performance of the present invention improves as the accuracy of the classifier increases. If the classifier identifies a frame as speech, the classifier assigns the frame an SNflag of 1. If the classifier identifies a frame as noise, the classifier assigns the frame an SNflag of 0. SNflag is a control value used to classify the frames.

Microprocessor 110 then determines additional speech waveform parameters of the speech signal according to the following equations:

$$Nfegy_n = fegy_n - Bfegy, n = 1,2,\ldots,N$$

The normalized frame energy, $Nfegy_n$, is the frame energy adjusted for noise. The bias frame energy, *Bfegy*, is an estimate of noise energy. It may be a theoretical or empirical number. It may also be measured, such as the noise in the first few frames of the speech acquisition window.

$$sum Nfegy_n = \sum_{j=1}^{n} Nfegy_j \, , n = 1,2, \ldots, N$$

In steps 220 through 235, microprocessor 110 determines whether the calculated energy centroid is within a speech region of the utterance. If a certain percent of frames before or after the energy centroid are noise frames, the energy centroid may not be within a speech region of the utterance. In this situation, microprocessor 110 will use the index of the peak energy as the starting point to determine the endpoints. The peak energy is usually expected to be within a speech region of the utterance. While the percent of noise frames surrounding the energy centroid has been chosen as the determining factor, it is understood that the percent of speech frames may be used as an alternative.

In step 220, microprocessor 110 determines whether the percent of noise frames in M1 frames preceding the energy centroid is greater than or equal to Valid1. While M1 may be any number of frames, M1 is preferably in the range of 5 to 20 frames. Valid1 is the percent of noise frames preceding the centroid and indicating the energy centroid is not within a speech region. While Valid1 could be any percent including 100 percent, Valid1 is preferably in the range of 70 to 100 percent. If the percent of noise frames in M1 frames preceding the energy centroid is greater than or equal to Valid1, then the frame index is set to be equal to the peak energy index, *epkindx*, in step 235. If the percent of noise frames in M1 frames preceding the energy centroid is less than Valid1, then the method proceeds to step 225.

In step 225, microprocessor 110 determines whether the percent of noise frames in M2 frames following the energy centroid is greater than or equal to Valid2. While M2 may be any number of frames, M2 is preferably in the range of 5 to 20 frames. Valid2 is the percent of noise frames following the centroid and indicating the energy centroid is not within a speech region. While Valid2 could be any percent including 100 percent, Valid1 is preferably in the range of 70 to 100 percent. If the percent of noise frames in M2 frames following the energy centroid is greater than or equal to Valid2, then the frame index is set to be equal to the peak energy index, *epkindx*, in step 235. If the percent of noise frames in M2 frames following the energy centroid is less than Valid2, then the frame index is set in step 230 to be equal to the index of the energy centroid, *icom*. With the frame index set in either step 230 or 235, the method proceeds to step 240.

are present. EMINP is a minimum percent of the total window energy. While EMINP may be any percent including 0 percent, EMINP is preferably within the range of 5 to15 percent. If the cumulative energy at STRTNP is greater than EMINP of the total window energy, then STRPNT is not an endpoint. The method proceeds to step 250, where microprocessor 110 decrements STRPNT by X frames. The method then continues to step 245.

If the cumulative energy at STRTNP is less than or equal to EMINP of the total window energy, then the current value of STRPNT is the start endpoint. The method proceeds to step 260, where the speech start index is equal to the current value for STRPNT. The method continues to step 265 for microprocessor 110 to determine the end endpoint.

In steps 265 through 285, microprocessor 110 determines the end endpoint of the speech utterance. Microprocessor 110 begins at the Frame Index, basically at a position within the speech region of the utterance, and analyzes the frames following the Frame Index to identify a potential end endpoint. When a potential end endpoint is identified, microprocessor 110 checks whether the cumulative frame energy at the potential end endpoint is greater than or equal to a percent of the total window energy. If the potential end endpoint is the end endpoint of the utterance, the cumulative frame energy at that frame should be almost all if not all of the total window energy. The cumulative frame energy at such frame indicates whether additional speech frames are present. In this manner, gaps and pauses in the utterance will not result in a erroneous end endpoint determination.

In step 265, microprocessor 110 sets ENDPNT equal to the Frame Index. ENDPNT is the frame being tested as the end endpoint. While ENDPNT is equal to the Frame Index initially, microprocessor 110 will increment ENDPNT until the end endpoint is found.

In step 270, microprocessor 110 determines whether the percent of noise frames in M4 frames following ENDPNT is greater than or equal to Test2. While M4 can be any number of frames, M4 is preferably in the range of 5 to 20 frames. Test2 is the percent of noise frames indicating ENDPNT is an endpoint. While Test2 could be any percent including 100 percent, Test2 is preferably in the range of 70 to 100 percent.

CLAIMS

1. A communication device capable of endpointing speech utterances, comprising:

at least one microprocessor having a speech/noise classifier,

wherein the at least one microprocessor analyzes a speech signal to determine speech waveform parameters within a speech acquisition window, wherein the speech waveform parameters include a cumulative frame energy, an energy centroid of the speech waveform, and a total window energy,

wherein the at least one microprocessor identifies a potential endpoint by analyzing frames in the speech acquisition window in relation to the energy centroid, and

wherein the at least one microprocessor validates the potential endpoint is an endpoint by comparing the cumulative frame energy at the potential endpoint to the total window energy;

a microphone for providing the speech signal to the at least one microprocessor; and

at least one communication output mechanism.

6. A method for endpointing speech utterances, wherein the speech utterances have a start endpoint and an end endpoint, comprising the steps of:

(a) analyzing a speech signal to determine speech waveform parameters within a speech acquisition window, wherein the speech waveform parameters include a cumulative frame energy, an energy centroid of the speech waveform, and a total window energy;

(b) identifying a potential start endpoint by analyzing at least one of noise and speech in frames in the speech acquisition window that precede the energy centroid; and

(c) validating the potential start endpoint is the start endpoint by comparing the cumulative frame energy at the potential start endpoint to the total window energy.

Application No: GB 0008337.8
Claims searched: 1 to 10

Examiner: John Donaldson
Date of search: 16 June 2000

Patents Act 1977
Search Report under Section 17

Databases searched:

UK Patent Office collections, including GB, EP, WO & US patent specifications, in:

UK Cl (Ed.R): G4R(RPS, RPW)

Int Cl (Ed.7): G10L 11/00, 11/02, 15/00, 15/04, 15/20

Other: Online:WPI, EPODOC, JAPIO

Documents considered to be relevant:

Category	Identity of document and relevant passage	Relevant to claims
A	GB 2090453 A (WESTERN ELECTRIC), see abstract	-

X	Document indicating lack of novelty or inventive step	A	Document indicating technological background and/or state of the art.
Y	Document indicating lack of inventive step if combined with one or more other documents of same category.	P	Document published on or after the declared priority date but before the filing date of this invention.
&	Member of the same patent family	E	Patent document published on or after, but with priority date earlier than, the filing date of this application.